Impact of class clustering in a multiclass FCFS queue with order-dependent service times

Bert Réveil, Dieter Claeys, Tom Maertens, Joris Walraevens, Herwig Bruneel

SMACS Research Group, TELIN Department, Ghent University
Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium
E-mail: {breveil, dclaeys, tmaerten, jw, hb}@telin.ugent.be

Abstract In multi-class queueing systems, customers of different classes can enter the system. When studying such systems, it is traditionally assumed that the different classes of customers occur randomly and independently in the arrival stream of customers in the system. This is often in contrast to the actual situation. Therefore, we study a multi-class system with so-called class clustering in the customer arrival stream, i.e., (Markovian) correlation occurs in the classes of consecutive customers. The system under investigation consists of one server that is able to serve two classes of customers. In addition, the service-time distribution of a customer depends on the equality or non-equality of its class with the class of the previous customer. This latter feature occurs frequently in practice. For instance, execution of the same task again can lead to both faster or slower processing times. The first case can occur when the execution of a different task entails resetting a machine, or loading new data, et cetera. The opposite situation appears, for instance, when execution of the same task requires postprocessing (such as cooling down or reinitialisation of a machine). We deduce the probability generating function (pgf) of the system content, from which we can extract various performance measures, among which the mean values of the system content and the customer delay. We demonstrate that class clustering has a tremendous impact on the system performance, which highlights the necessity to include it in the performance assessment of any system in which it occurs.

keywords: queueing, order-dependent service times, class clustering

1 Introduction

In multi-class queueing systems, customers of different classes can enter the system. Customers of distinct classes, for instance, correspond to data packets requiring transmission over different egress lines, vehicles that are heading to other destinations, data packets with different priorities, jobs with other execution times, people desiring distinct kinds of service at a call center, patients in the waiting room with different complaints, et cetera. As multi-class systems appear frequently in practice, they have attracted a lot of attention in the literature. However, in most studies it is standard to assume that the different classes of customers occur randomly and independently in the arrival stream of customers into the system (see e.g. [1, 3, 8, 9, 14, 16, 22, 23] and references therein), which is often in contrast to the actual situation. In reality, there is often some degree of *interclass correlation* or *class clustering*. In some cases, for instance, customers of the same class have a tendency to arrive "back to back". To see that, one can think of a network router that transmits data from and towards various communicating processes that are running in the network. Within certain time frames of its service, it is not unlikely that the router will transfer subsequent data packets that all originate from the same process, i.e., packets of the same class are processed in clusters.

In some discrete-time multi-class queueing models correlation exists between the numbers of arrivals of different classes in the same slot, but the numbers of arrivals of each class during consecutive slots are

independent and identically distributed [15, 20, 19, 24]. Hence, in [15, 20, 19, 24], correlation is defined slot-based, whereas our definition is customer-based, i.e., we let the classes of consecutive customers in the arrival process be dependent, even across slot boundaries. As a result, class clustering is not adequately modelled in [15, 20, 19, 24], whereas the opposite holds in this paper. We believe that the arrival process considered in this paper may be more suitable for particular applications, such as in manufacturing systems or in the scheduling of computing jobs.

In order to examine the impact of class clustering, we recently studied it in various multi-class queueing systems [6, 7, 18, 17, 5]. In [18], a continuous-time system was considered with two classes of customers that each have their own dedicated server and that are accommodated in one common queue. Such a system is useful in the modeling of e.g. traffic junctions in road traffic or input queues in packet switches. In [6, 7], we evaluated the analogous discrete-time variant of the above system, as this variant models the behaviour of telecommunications systems more closely. In [10], it was shown that the results of [6, 7] can also be applied to study data clustering on in-order processing systems. A system with one server, two queues, and customers with two priority classes has been examined in [17]. The latter system can be applied to model the distinction between data and real-time packets in telecommunications systems. Finally, in [5], a system was evaluated with one server, one queue, and two classes of customers that require distinct service times. The latter feature can, for instance, model the time that patients visit the doctor, as this depends on the type of complaints they have. In all the examined systems, we have demonstrated that class clustering can have a huge influence on the overall system performance.

The current paper is motivated by the observation that, in practice, it can occur that the service time of a customer does not necessarily depend on its own class, but rather on the equality or non-equality of its class with the class of the previous customer. For instance, consider software programs that run on the same server and that require lots of program-specific data. If program P was executed during the previous run, the necessary data will have been loaded into the cache memory, and hence execution of program P in the current run will be faster. As a second example, we mention a specialized printing house that delivers print work in several different formats. If subsequent customers ask for the same format, the machine can start printing immediately. In the opposite case, some mechanical parts of the printing machine may have to be reset, before printing can be initiated.

Based on the conclusions of the cited papers above, we have reason to believe that class clustering will also have an important impact on the behaviour of systems where the service time of a customer depends on the equality or non-equality of its class with the class of the previous customer. Hence, when studying the performance of such systems, class clustering should be incorporated in order to obtain realistic results and to be able to quantify the impact of it. This forms the topic of this paper. More specifically, we describe the system under investigation in detail in section 2. It is a system where two classes of customers are placed in a common FCFS queue, where class clustering is included and whereby the service time of a customer depends on the equality of its class with the class of the previous customer. In section 3, we analyse the system behaviour and we derive an expression for the pgf of the number of customers in the system, hereafter referred to as the *system content*, both at customer departure times as at random slot boundaries. Some special system cases are discussed in section 4. Finally, the influence of class clustering is investigated in section 5, and some conclusions are drawn in section 6.

2 System description

We study a discrete-time queueing system with an infinite waiting room, one server, and two classes of customers, named 1 and 2. As in all discrete-time systems, the time axis is divided into fixed-length time

intervals that are referred to as *slots* in the sequel. New customers can arrive in the system at any given (continuous) point on the time axis, but customer service times can only start and end at slot boundaries. Customers are served according to a so-called *global FCFS* service discipline, meaning that they are served in order of their arrival, regardless of the class they belong to.

The arrival process of new customers is characterized in two steps. First, the total (aggregated) number of customer arrivals in subsequent slots is represented by a sequence of independent and identically distributed (i.i.d.) nonnegative discrete random variables with common probability mass function (pmf) $e(n)$ and common probability generating function (pgf) $E(z)$:

$$e(n) \triangleq \text{Prob}[n \text{ arrivals in one slot}] \ , \ n \geq 0 \ ,$$
$$E(z) \triangleq \sum_{n=0}^{\infty} e(n) z^n .$$

The *(total) mean arrival rate*, that is the (total) mean number of arrivals per slot, is then given by

$$\lambda \triangleq E'(1) \ .$$

Secondly, the occurrence of class 1 and class 2 customers within the total arrival stream is governed by a customer-class correlation model. This implies that we account for the possibility of *interclass correlation*, or *class clustering* in the arrival process. Customers of any given class may (or may not) have a tendency to "arrive back-to-back". Consequently, the classes of two consecutive customers may be non-independent. In this study, we consider a discrete-time first-order Markov chain to model correlation between the classes of two consecutive customers. The transition probabilities of the Markov chain are defined as (see Fig. 1)

$$\alpha \triangleq \text{Prob}[t_{k+1} = 1 | t_k = 1]; \beta \triangleq \text{Prob}[t_{k+1} = 2 | t_k = 2] \ , \tag{1}$$

with t_k the class of customer k. The steady-state probabilities of finding the Markov chain in state 1 respec-

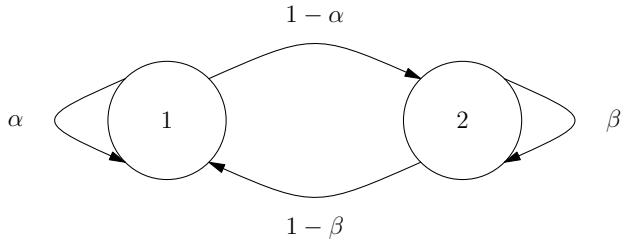

Fig. 1. Two-state Markov chain of the customer classes.

tively 2 are equal to [11, 13]

$$\pi_1 \triangleq \lim_{k \to \infty} \text{Prob}[t_k = 1] = \frac{1-\beta}{2-\alpha-\beta} \ , \ \pi_2 \triangleq \lim_{k \to \infty} \text{Prob}[t_k = 2] = \frac{1-\alpha}{2-\alpha-\beta} \ .$$

They can be interpreted as the fractions of class 1 and class 2 customers in the arrival stream. The steady-state correlation coefficient γ ($-1 \leq \gamma \leq 1$) of the Markov chain, referred to as the *interclass correlation* in

the sequel, is defined as

$$\gamma \triangleq \lim_{k\to\infty} \frac{\mathrm{E}\left[t_k t_{k+1}\right] - \mathrm{E}\left[t_k\right]\mathrm{E}\left[t_{k+1}\right]}{\sqrt{\mathrm{var}[t_k]\ \mathrm{var}[t_{k+1}]}} \ = \ \alpha + \beta - 1 \ .$$

It represents the amount of correlation between the classes of two consecutive customers in the arrival stream (in the steady-state). Positive values of γ correspond to situations in which at least one customer class has a tendency to cluster. Negative values of γ typically imply (strongly) alternating customer class arrivals. In case $\gamma = 0$, and consequently $\alpha = 1 - \beta$, the classes of consecutive customers are independent, and consequently, uncorrelated. This corresponds with the situation that is traditionally (implicitly) assumed in literature.

The service process of the customers is characterized by means of two possible *service-time* distributions. The *service time*, or *service requirement* of a customer indicates the number of slots that is needed to fully serve that customer. Concretely, we assume that the service time of a customer depends on its own class and on the class of the previous customer. If both classes are the same, the pmf of the service time is given by

$$a(n) \triangleq \mathrm{Prob}[n \text{ slots needed to serve customer } k | t_k = t_{k-1}] \ , \ n \geq 1 \ ,$$

otherwise, the pmf is given by

$$b(n) \triangleq \mathrm{Prob}[n \text{ slots needed to serve customer } k | t_k \neq t_{k-1}] \ , \ n \geq 1 \ .$$

The corresponding pgfs are denoted by

$$A(z) \triangleq \sum_{n=1}^{\infty} a(n) z^n \ , \ B(z) \triangleq \sum_{n=1}^{\infty} b(n) z^n \ .$$

The mean service times for customers following a same-class customer or a customer of the opposite class are given by

$$\mu_A \triangleq A'(1) \ , \ \mu_B \triangleq B'(1) \ .$$

3 System analysis

In this section, we first present an analysis of the total number of customers in the system at customer departure times. An expression is derived for the pgf of this number (under steady-state conditions) and a method is described that can be used to determine the two remaining unknowns in the expression. Next, we also provide an expression for the pgf of the system content and the average system content at random slot boundaries. All given derivations are valid for arbitrary choices of the pgfs $E(z)$, $A(z)$ and $B(z)$, and for arbitrary α and β values.

3.1 System equations at customer departure times

In this subsection, we establish system equations that capture the behaviour of the system content at customer departure times. To that end, we introduce the variable u_k, the total number of customers in the system immediately after the service completion of the k-th customer. Due to the assumptions presented in section 2, the sequence of couples $\{(t_k,u_k)\}$ constitutes a discrete-time Markov chain. As was described

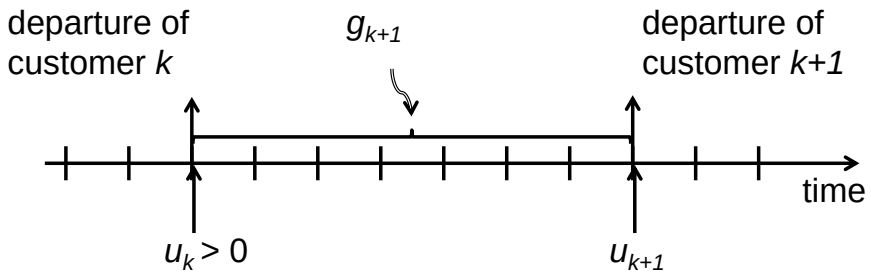

Fig. 2. Relationship between u_k and u_{k+1} when $u_k > 0$.

above, the state transitions for the sequence $\{t_k\}$ are governed by Equation (1). For the quantities $\{u_k\}$, we obtain two recursive equations that cover the complete set of situations that is depicted in figures 2 and 3:

$$\begin{aligned} u_{k+1} &= u_k - 1 + g_{k+1} \text{ , if } u_k > 0 \text{ ,} \\ u_{k+1} &= h_{k+1} \text{ , if } u_k = 0 \text{ .} \end{aligned} \tag{2}$$

In these equations, g_{k+1} stands for the (total) number of arrivals in the system during the service time of customer $k+1$. The quantity h_{k+1} is defined as

$$h_{k+1} \triangleq g_{k+1} + f_{k+1} \text{ ,}$$

with f_{k+1} indicating the number of customer arrivals in the arrival slot of customer $k+1$, but *after* customer $k+1$ (in case customer $k+1$ enters an empty system).

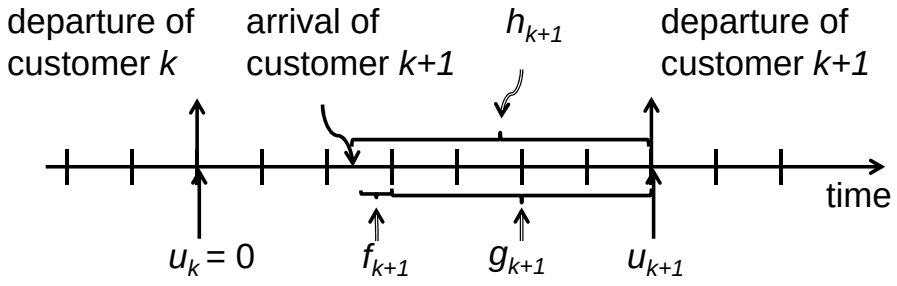

Fig. 3. Relationship between u_k and u_{k+1} when $u_k = 0$.

As f_{k+1} represents the number of additional arrivals in a slot with at least one arrival, its pmf $f(n)$ and pgf $F(z)$ can be found as

$$f(n) \triangleq \text{Prob}[n \text{ additional arrivals}|\text{at least 1 arrival}] = \frac{e(n+1)}{1-E(0)} \ , \ n \geq 0 \ ,$$
$$F(z) \triangleq E[z^{f_{k+1}}] \ = \ \sum_{n=0}^{\infty} f(n) z^n \ = \ \frac{E(z)-E(0)}{z[1-E(0)]} \ , \tag{3}$$

irrespective of whether customer $k+1$ is of the same or different class as customer k. For the pgfs of the quantities g_{k+1} and h_{k+1}, the equality of the customer classes of two consecutive customers does make a difference. Taking into account that we are considering an i.i.d. aggregated arrival process, which implies that f_{k+1} and g_{k+1} are mutually independent, we find that

$$G_A(z) \triangleq \mathrm{E}\left[z^{g_{k+1}}|t_{k+1}=t_k\right] = A(E(z)) \ , \ H_A(z) \triangleq \mathrm{E}\left[z^{h_{k+1}}|t_{k+1}=t_k\right] = F(z)A(E(z)) \ ,$$
$$G_B(z) \triangleq \mathrm{E}\left[z^{g_{k+1}}|t_{k+1}\neq t_k\right] = B(E(z)) \ , \ H_B(z) \triangleq \mathrm{E}\left[z^{h_{k+1}}|t_{k+1}\neq t_k\right] = F(z)B(E(z)) \ .$$

3.2 System content at customer departure times

One of our intentions is to provide expressions for the performance measures of the queueing system under so-called steady-state conditions. This means that we assume that, after an initial transient phase, our system is operating in a stable regime. It is well-known [4, 21] that for any work-conserving queueing system stability is guaranteed as soon as the average amount of work entering the system per slot (often referred to as the *work load* ρ) is strictly less than the amount of work that can be delivered by the server per slot. In our model, considering a single server without interruptions, the stability condition thus boils down to

$$\rho \triangleq \lambda \mathrm{E}\left[c\right] < 1 \ ,$$

with c the service time of an arbitrary customer. Using the law of the total expectation, $\mathrm{E}\left[c\right]$ can be expanded, yielding

$$\mathrm{E}\left[c\right] = \pi_A \mu_A + (1-\pi_A)\mu_B \ , \tag{4}$$

where π_A denotes the steady-state probability that two consecutive customers belong to the same class:

$$\pi_A \triangleq \lim_{k\to\infty} \text{Prob}[t_k = t_{k-1}]$$
$$= \alpha\pi_1 + \beta\pi_2 = \frac{1-\beta}{2-\alpha-\beta}\alpha + \frac{1-\alpha}{2-\alpha-\beta}\beta \ .$$

If we rework Equation (4), substituting α and β in terms of γ, π_1, and π_2, another, more interesting expression for ρ can be found that links the work load directly to the amount of interclass correlation in the arrival process:

$$\rho = \lambda[\mu_A + 2(1-\gamma)(\mu_B - \mu_A)\pi_1\pi_2] \ . \tag{5}$$

As could have been anticipated, we find that in case of ultimate positive customer class correlation ($\gamma = 1$), the work load reduces to $\lambda\mu_A$. Moreover, this also holds for single-class systems where α or β are equal to 1,

because in that case either π_1 or π_2 equals 0. It is also true if $\mu_A = \mu_B$. If γ equals -1, i.e., if the customer class changes with every customer arrival, π_1 and π_2 are both equal to 0.5, and the work load reduces to $\lambda\mu_B$, also as expected.

Assuming that the stability condition is met, we define joint steady-state probabilities for the Markov chain $\{(t_k,\, u_k)\}$ as

$$p_1(i) \triangleq \lim_{k\to\infty} \text{Prob}[t_k = 1,\, u_k = i] \;,\; p_2(i) \triangleq \lim_{k\to\infty} \text{Prob}[t_k = 2,\, u_k = i] \;,$$

for all $i \geq 0$. The corresponding partial pgfs are equal to

$$P_1(z) \triangleq \sum_{i=0}^{\infty} p_1(i) z^i \;,\; P_2(z) \triangleq \sum_{i=0}^{\infty} p_2(i) z^i \;,$$

while the steady-state pgf $P(z)$ of the total system content at customer departure times is given by

$$P(z) = P_1(z) + P_2(z) \;. \tag{6}$$

Relying on the balance equations of the Markov chain, it is now possible to establish two linearly independent equations for the partial pgfs $P_1(z)$ and $P_2(z)$. For customers of class 1, we get

$$\begin{aligned}
p_1(j) &= \lim_{k\to\infty} \text{Prob}[t_{k+1} = 1,\, u_{k+1} = j] \\
&= \sum_{i=0}^{\infty} \lim_{k\to\infty} \text{Prob}[t_k = 1,\, u_k = i] \text{Prob}[t_{k+1} = 1,\, u_{k+1} = j | t_k = 1,\, u_k = i] \\
&\quad + \sum_{i=0}^{\infty} \lim_{k\to\infty} \text{Prob}[t_k = 2,\, u_k = i] \text{Prob}[t_{k+1} = 1,\, u_{k+1} = j | t_k = 2,\, u_k = i] \qquad (7) \\
&= \alpha \sum_{i=0}^{\infty} p_1(i) \lim_{k\to\infty} \text{Prob}[u_{k+1} = j | u_k = i,\, t_k = 1,\, t_{k+1} = 1] \\
&\quad + (1-\beta) \sum_{i=0}^{\infty} p_2(i) \lim_{k\to\infty} \text{Prob}[u_{k+1} = j | u_k = i,\, t_k = 2,\, t_{k+1} = 1] \;.
\end{aligned}$$

Taking the z-transform of (7) yields:

$$\begin{aligned}
P_1(z) &= \alpha \sum_{i=0}^{\infty} p_1(i) \lim_{k\to\infty} \text{E}\left[z^{u_{k+1}} | u_k = i,\, t_k = 1,\, t_{k+1} = 1\right] \qquad (8) \\
&\quad + (1-\beta) \sum_{i=0}^{\infty} p_2(i) \lim_{k\to\infty} \text{E}\left[z^{u_{k+1}} | u_k = i,\, t_k = 2,\, t_{k+1} = 1\right] .
\end{aligned}$$

The expectations in the above equations can be further developed using the system equations in (2):

$$\begin{aligned}
\lim_{k\to\infty} \mathrm{E}\left[z^{u_{k+1}}|u_k = i,\, t_k = 1,\, t_{k+1} = 1\right] &= \lim_{k\to\infty} \mathrm{E}\left[z^{i-1+g_{k+1}}|t_k = 1,\, t_{k+1} = 1\right] \\
&= z^{i-1}G_A(z) \text{ , for all } i \geq 1 \text{ ,} \qquad (9)\\
\lim_{k\to\infty} \mathrm{E}\left[z^{u_{k+1}}|u_k = 0,\, t_k = 1,\, t_{k+1} = 1\right] &= \lim_{k\to\infty} \mathrm{E}\left[z^{h_{k+1}}|t_k = 1,\, t_{k+1} = 1\right] \\
&= H_A(z) \text{ , for } i = 0 \text{ ,} \qquad (10)\\
\lim_{k\to\infty} \mathrm{E}\left[z^{u_{k+1}}|u_k = i,\, t_k = 2,\, t_{k+1} = 1\right] &= \lim_{k\to\infty} \mathrm{E}\left[z^{i-1+g_{k+1}}|t_k = 2,\, t_{k+1} = 1\right] \\
&= z^{i-1}G_B(z) \text{ , for all } i \geq 1 \text{ ,} \qquad (11)\\
\lim_{k\to\infty} \mathrm{E}\left[z^{u_{k+1}}|u_k = 0,\, t_k = 2,\, t_{k+1} = 1\right] &= \lim_{k\to\infty} \mathrm{E}\left[z^{h_{k+1}}|t_k = 2,\, t_{k+1} = 1\right] \\
&= H_B(z) \text{ , for } i = 0 \text{ .} \qquad (12)
\end{aligned}$$

Substitution of (9), (10), (11) and (12) in equation (8) finally leads to a first linear equation between $P_1(z)$ and $P_2(z)$:

$$\begin{aligned}
P_1(z) = \alpha P_1(0)H_A(z) + \frac{\alpha G_A(z)}{z}(P_1(z) - P_1(0)) \\
+(1-\beta)P_2(0)H_B(z) + \frac{(1-\beta)G_B(z)}{z}(P_2(z) - P_2(0)) \text{ ,}
\end{aligned}$$

or, reworked in terms of the arrival pgf $E(z)$ and service-time pgfs $A(z)$ and $B(z)$:

$$\begin{aligned}
(z - \alpha A(E(z)))P_1(z) = (1-\beta)B(E(z))P_2(z) + \alpha A(E(z))(zF(z)-1)P_1(0) \\
+(1-\beta)B(E(z))(zF(z)-1)P_2(0) \text{ .} \qquad (13)
\end{aligned}$$

Along the same lines, a second linear equation can be found starting from the balance equations for class 2 customers:

$$\begin{aligned}
(z - \beta A(E(z)))P_2(z) = (1-\alpha)B(E(z))P_1(z) + \beta A(E(z))(zF(z)-1)P_2(0) \\
+(1-\alpha)B(E(z))(zF(z)-1)P_1(0) \text{ .} \qquad (14)
\end{aligned}$$

Equations (13) and (14) can be solved for the unknown partial pgfs $P_1(z)$ and $P_2(z)$. Using the results and Equation (3) to expand Equation (6), we then obtain a first expression for the pgf $P(z)$:

$$P(z) = \frac{P(0)(E(z)-1)}{1-E(0)} \frac{z(p_A A(E(z)) + p_B B(E(z))) - \alpha\beta A(E(z))^2 + (1-\alpha)(1-\beta)B(E(z))^2}{z^2 - z(\alpha+\beta)A(E(z)) + \alpha\beta A(E(z))^2 - (1-\alpha)(1-\beta)B(E(z))^2} \text{ ,} \qquad (15)$$

where we have introduced the following definitions for the quantities p_A and p_B:

$$p_A \triangleq \frac{\alpha P_1(0) + \beta P_2(0)}{P(0)} \text{ , } p_B \triangleq \frac{(1-\alpha)P_1(0) + (1-\beta)P_2(0)}{P(0)} \text{ .} \qquad (16)$$

Given that $P(0) = P_1(0) + P_2(0)$, these quantities can be seen as conditional probabilities

$$p_A = \lim_{k\to\infty} \mathrm{Prob}[t_{k+1} = t_k | u_k = 0] \text{ , } p_B = \lim_{k\to\infty} \mathrm{Prob}[t_{k+1} \neq t_k | u_k = 0] \text{ ,}$$

the conditional probabilities that a new customer entering an empty system (in regime) belongs to the same or the opposite class respectively as the last customer that was served by the system.

Expression (15) still contains three unknowns that need to be determined: $P(0)$, p_A and p_B. The probability $P(0)$ can be found by imposing the normalization condition on the pgf $P(z)$, i.e. $P(1) = 1$. Using de l'Hôpital's rule to solve the equation, we obtain that

$$P(0) = \frac{(1 - E(0))(1 - \rho)}{\lambda} \ . \tag{17}$$

In order to derive expressions for p_A and p_B, two linear equations in p_A and p_B are established. The first one simply states that

$$p_A + p_B = 1 \ . \tag{18}$$

To obtain the second equation, we will prove that the denominator of $P(z)$ has exactly two zeroes inside the closed complex unit disk $\{z \in \mathbb{C} : |z| \leq 1\}$. Due to the analytic property of pgfs inside the unit disk, those zeroes, one of which is equal to 1, must also be zeroes of $P(z)$'s numerator. The combination of the zero distinct from 1 and the analytic property of pgfs will produce the second equation.

In what follows, we prove that the denominator of $P(z)$ has 2 zeroes inside the closed complex unit disk. In many papers, the denominator equals $z^n - X(z)$, with n a positive integer, $X(z)$ a pgf and $X^{'}(1) < n$ (this condition generally is equivalent with the stability condition). As it is not difficult to prove by means of Rouché's theorem ([2], [21]) that the (more general) function $z^n - KX(z)$ (with $K \leq 1$) has exactly n zeroes in the closed complex unit disk, application of Rouché's theorem is straightforward in most papers, and is therefore often omitted. Unfortunately, in the current paper, the denominator of $P(z)$ (see equation (15)) has a more intricate form, and it is therefore more difficult to prove that it has 2 zeroes inside the closed complex unit disk. We present the proof in the following theorem:

Theorem 1. *The denominator of $P(z)$ (equation (15)) has two zeroes in the closed complex unit disk.*

Proof. The germ of our approach is to structure the denominator of $P(z)$ as $f_1(z)f_2(z) - g(z)$, with

$$f_1(z) \triangleq z - \alpha A(E(z)) \ ,$$

$$f_2(z) \triangleq z - \beta A(E(z)) \ ,$$

$$g(z) \triangleq (1 - \alpha)(1 - \beta)B(E(z))^2 \ .$$

Note that these are all analytic functions inside the closed complex unit disk, and that $f_1(z)$ and $f_2(z)$ belong to the class of functions for which it is straightforward to prove by means of Rouché's theorem that they contain one zero inside the closed complex unit disk. As a result, $f_1(z)f_2(z)$ contains 2 zeroes inside the closed complex unit disk. We now prove that

$$|f_1(z)f_2(z)| > |g(z)| \ ,$$

on the contour $|z| = 1 + \epsilon$, with ϵ a (very) small increment larger than 0. We can write that

$$\begin{aligned}\left|\frac{g(z)}{f_1(z)f_2(z)}\right| &= |B(E(z))|^2 \left|\frac{1-\alpha}{z-\alpha A(E(z))}\right| \left|\frac{1-\beta}{z-\beta A(E(z))}\right| \\ &= |B(E(z))|^2 \left|\frac{G_1\left(\frac{A(E(z))}{z}\right)}{z}\right| \left|\frac{G_2\left(\frac{A(E(z))}{z}\right)}{z}\right| \ ,\end{aligned}$$

with

$$G_1(z) \triangleq \frac{1-\alpha}{1-\alpha z} \ ,$$

and

$$G_2(z) \triangleq \frac{1-\beta}{1-\beta z} \ ,$$

pgfs of geometric distributions with parameters α and β respectively. Next, let us consider values of z satisfying $|z| = 1 + \epsilon$. First, we have

$$\begin{aligned}|B(E(z))|^2 &\leq B(E(1+\epsilon)))^2 \\ &= \left(1 + \epsilon\mu_B\lambda + O(\epsilon^2)\right)^2 \\ &= 1 + \epsilon 2\mu_B\lambda + O(\epsilon^2) \ .\end{aligned}$$

Next, it holds that

$$\begin{aligned}\left|\frac{G_1\left(\frac{A(E(z))}{z}\right)}{z}\right| &\leq \frac{G_1\left(\frac{A(E(1+\epsilon))}{1+\epsilon}\right)}{1+\epsilon} \\ &= \frac{G_1\left(\frac{1+\epsilon\mu_A\lambda+O(\epsilon^2)}{1+\epsilon}\right)}{1+\epsilon} \\ &= 1 + \epsilon\left[-1 + (-1+\mu_A\lambda)\frac{\alpha}{1-\alpha}\right] + O(\epsilon^2) \ ,\end{aligned}$$

where we have invoked $G_1'(1) = \alpha/(1-\alpha)$ and $1/(1+\epsilon) = 1 - \epsilon + O(\epsilon^2)$. Analogously, we find

$$\left|\frac{G_2\left(\frac{A(E(z)))}{z}\right)}{z}\right| \leq 1 + \epsilon\left[-1 + (-1+\mu_A\lambda)\frac{\beta}{1-\beta}\right] + O(\epsilon^2) \ .$$

Combining the above results, we obtain

$$\left|\frac{g(z)}{f_1(z)f_2(z)}\right| \leq 1 + \epsilon\left[2\mu_B\lambda - 1 + \frac{\alpha}{1-\alpha}[-1+\mu_A\lambda] - 1 + \frac{\beta}{1-\beta}[-1+\mu_A\lambda]\right] + O(\epsilon^2) \ .$$

The coefficient corresponding to ϵ can be rewritten as

$$\frac{\lambda}{(1-\alpha)(1-\beta)}\left[\{\alpha(1-\beta)+\beta(1-\alpha)\}\mu_A+2(1-\alpha)(1-\beta)\mu_B\right]-\frac{1}{(1-\alpha)(1-\beta)}\left[2-\alpha-\beta\right] \ .$$

As the stability condition $\rho<1$ can be expressed as

$$\lambda\left[\{\alpha(1-\beta)+\beta(1-\alpha)\}\mu_A+2(1-\alpha)(1-\beta)\mu_B\right]<2-\alpha-\beta \ ,$$

the coefficient corresponding to ϵ is strictly smaller than 0. Hence, on the contour $|z|=1+\epsilon$, it holds that $|f_1(z)f_2(z)|>|g(z)|$. On account of Rouché's theorem, $f_1(z)f_2(z)-g(z)$, i.e., the denominator of $P(z)$, has 2 zeroes inside $\{z\in\mathbb{C}:|z|\leq 1+\epsilon\}$. Letting $\epsilon\to 0$ concludes the proof. □

For $z=1$, given its factor $(E(z)-1)$, the numerator of $P(z)$ clearly vanishes. For the second zero however, called $\hat{z}$ from here on, the other factor in the numerator should equal 0, which yields a linear equation for p_A and p_B. Solving this equation, in combination with Equation (18), we find that p_A and p_B can be determined as

$$p_A=\frac{(\alpha+\beta)A(E(\hat{z}))-B(E(\hat{z}))-\hat{z}}{A(E(\hat{z}))-B(E(\hat{z}))} \ , \ p_B=\frac{(1-\alpha-\beta)A(E(\hat{z}))+\hat{z}}{A(E(\hat{z}))-B(E(\hat{z}))} \ .$$

Once the zero $\hat{z}$ is computed (numerically), p_A and p_B are fixed, and as such, so is $P(z)$.

3.3 System content at random slot boundaries

From earlier research [4], it follows that for all discrete-time queueing systems that incorporate one single server and generally independent customer arrivals from slot to slot (with pgf $E(z)$), a fairly simple relationship holds between the pgf $P(z)$ of the system content at customer departure times and the pgf $U(z)$ of the system content at random slot boundaries, regardless of the exact characteristics of the service process and the intra-slot details of the arrival process (e.g., single or batch arrivals, when do customers arrive within the slot, etc.). This relationship is

$$P(z)=\frac{E(z)-1}{\lambda(z-1)}U(z) \ . \tag{19}$$

As the examined model belongs to the class of systems that is described above, relationship (19) in combination with Equations (15) and (17) leads to the following expression for the pgf of the system content at random slot boundaries:

$$U(z)=\frac{(1-\rho)(z-1)[z(p_AA(E(z))+p_BB(E(z)))-\alpha\beta A(E(z))^2+(1-\alpha)(1-\beta)B(E(z))^2]}{z^2-z(\alpha+\beta)A(E(z))+\alpha\beta A(E(z))^2-(1-\alpha)(1-\beta)B(E(z))^2} \ . \tag{20}$$

From this expression, various interesting performance measures can be derived, one of which is of course the mean system content $\mathrm{E}[u]$ at random slot boundaries. The latter can be determined as $\mathrm{E}[u]=U'(1)$. After long and tedious calculations, we find that

$$\begin{aligned}\mathrm{E}[u]=\rho&+\frac{\lambda^2C''(1)+E''(1)C'(1)}{2(1-\rho)}+2\lambda(\mu_A-\mu_B)\pi_1\pi_2+\frac{p_B\lambda(\mu_B-\mu_A)}{1-\gamma}\\&+\frac{(1-\gamma)\lambda^2(\mu_B-\mu_A)^2\pi_1\pi_2(1-4\pi_1\pi_2)}{1-\rho} \ ,\end{aligned} \tag{21}$$

with $C'(1)$ and $C''(1)$ the first two derivatives at $z = 1$ of the pgf $C(z)$ of the service time of an arbitrary customer:

$$C(z) = \pi_A A(z) + (1 - \pi_A)B(z) \ .$$

In Equation (21), the first term ρ accounts for the average server content, or the mean number of customers in service. The last four terms cover the mean queue occupancy, meaning the average number of customers that are actually waiting to be served.

Higher-order moments of the system content at random slot boundaries can be obtained by computing higher-order derivatives of the pgf $U(z)$. By means of Little's law (for discrete-time queues) [12], one can determine the average *delay* (system time) of an arbitrary customer as $\mathrm{E}[d] = \mathrm{E}[u]/\lambda$. The mean *waiting time* $\mathrm{E}[w]$ is obtained as $\mathrm{E}[w] = \mathrm{E}[d] - \mathrm{E}[c]$, where $\mathrm{E}[c]$ was defined in (4). We obtain

$$\begin{aligned}\mathrm{E}[w] = &\frac{\lambda^2 C''(1) + E''(1)C'(1)}{2\lambda(1-\rho)} + 2(\mu_A - \mu_B)\pi_1\pi_2 + \frac{p_B(\mu_B - \mu_A)}{1-\gamma} \\ &+ \frac{(1-\gamma)\lambda(\mu_B - \mu_A)^2\pi_1\pi_2(1 - 4\pi_1\pi_2)}{1-\rho} \ .\end{aligned}$$

4 Four special cases

In this section, we study a few special examples of our general system model.

4.1 Identical service-time distributions: $A(z) = B(z)$

A first special model occurs when the service-time distribution of customers does not depend on the equality of the subsequent customer classes any more, i.e., $A(z) = B(z)$. In that case, we expect to find the behaviour of a single-server system accommodating only one class of customers with service-time pgf $A(z)$. After rewriting the expression for $U(z)$, replacing $B(z)$ by $A(z)$, we indeed find the well-known [4, 21] pgf of the system content at random slot boundaries of such a model:

$$\begin{aligned}U(z) &= \frac{(1-\rho)(z-1)(zA(E(z)) + (1-\alpha-\beta)A(E(z))^2)}{z^2 - z(\alpha+\beta)A(E(z)) + (1-\alpha-\beta)A(E(z))^2} \\ &= \frac{(1-\rho)(z-1)A(E(z))(z + (1-\alpha-\beta)A(E(z)))}{(z - A(E(z)))(z + (1-\alpha-\beta)A(E(z)))} \\ &= \frac{(1-\rho)(z-1)A(E(z))}{z - A(E(z))} \ .\end{aligned} \tag{22}$$

4.2 Single-class system: $\alpha = 1$, or $\beta = 1$

A second special model is one where in the steady-state regime only one class of customers enters the system. This model corresponds to either α or β being 1. In that case, p_A becomes 1, while $p_B = 0$ (or, vice versa). Considering a model where $\alpha = 1$ and β takes on any arbitrary value, the expression for the pgf of the

system content at random slot boundaries reduces to

$$\begin{aligned} U(z) &= \frac{(1-\rho)(z-1)(zA(E(z)) - \beta A(E(z))^2)}{z^2 - z(1+\beta)A(E(z)) + \beta A(E(z))^2} \\ &= \frac{(1-\rho)(z-1)A(E(z))(z-\beta A(E(z)))}{(z-A(E(z)))(z-\beta A(E(z)))} \\ &= \frac{(1-\rho)(z-1)A(E(z))}{z-A(E(z))} \ , \end{aligned} \tag{23}$$

which is again what could be expected.

4.3 Service times independent of exact customer classes: $\alpha = \beta$

A third special system case appears when $\alpha = \beta$. Then, the probability of encountering an arrival of the same or the opposite class as the previous arrival becomes independent of the precise class of the previous customer. This implies that p_A and p_B are equal to α and $1-\alpha$ respectively (see Equation (16)). Reworking Equation (20) yields

$$\begin{aligned} U(z) &= \frac{(1-\rho)(z-1)[z(\alpha A(E(z)) + (1-\alpha)B(E(z))) - (\alpha A(E(z)))^2 + ((1-\alpha)B(E(z)))^2]}{z^2 - 2z\alpha A(E(z)) + (\alpha A(E(z)))^2 - ((1-\alpha)B(E(z)))^2} \\ &= \frac{(1-\rho)(z-1)[\alpha A(E(z)) + (1-\alpha)B(E(z))][z - (\alpha A(E(z)) - (1-\alpha)B(E(z)))]}{(z-\alpha A(E(z)))^2 - ((1-\alpha)B(E(z)))^2} \\ &= \frac{(1-\rho)(z-1)[\alpha A(E(z)) + (1-\alpha)B(E(z))][z - \alpha A(E(z)) + (1-\alpha)B(E(z))]}{[z - (\alpha A(E(z)) + (1-\alpha)B(E(z))][z - \alpha A(E(z)) + (1-\alpha)B(E(z))]} \\ &= \frac{(1-\rho)(z-1)C(E(z))}{z-C(E(z))} \ , \end{aligned} \tag{24}$$

with $C(z)$, pgf of the service time of an arbitrary customer, now equal to:

$$C(z) = \alpha A(z) + (1-\alpha)B(z) \ . \tag{25}$$

Note that expression (25) can also be written as

$$C(z) = \frac{\frac{\alpha}{1-\alpha}A(z) + B(z)}{\frac{\alpha}{1-\alpha} + 1} = \frac{A(z) + \frac{1-\alpha}{\alpha}B(z)}{1 + \frac{1-\alpha}{\alpha}} \ ,$$

where $\alpha/(1-\alpha)$ represents the average number of consecutive customers of the same class, and $\frac{1-\alpha}{\alpha}$ the average number of consecutive customers of different (i.e., alternating) classes.

4.4 No interclass correlation: $\gamma = 0$

A fourth interesting case occurs when the interclass correlation γ is equal to 0 (and consequently $\alpha = 1-\beta$). In that case, the probability of encountering a class 1 or a class 2 customer becomes independent of the class of the previous customer.

It turns out that this independence does not result in a special expression for the pgf of the system content at random slot boundaries, except when $\alpha = \beta = \frac{1}{2}$, which is a particular case of the set of systems that

was discussed above (Section 4.3). This is owed to the fact that, although $\gamma = 0$, in general the probability of observing subsequent customers of the same or the opposite class still depends on the specific class of the previous customer. Hence, the service time of customer $k+1$ is not independent of the service time of customer k, which is a necessary condition to find an expression of the form found in (22), (23) and (24). Subsequent service times being non-independent can, for instance, be observed when one considers the non-equality of the following two probabilities for the proposed system case ($\gamma = 0$):

$$\text{Prob}[t_{k+1} \neq t_k | t_k \neq t_{k-1}] \text{ , and } \text{Prob}[t_{k+1} \neq t_k] \text{ .}$$

While the latter probability equals $2\alpha(1-\alpha)$:

$$\begin{aligned}\text{Prob}[t_{k+1} \neq t_k] &= \text{Prob}[t_{k+1} = 1 | t_k = 2]\text{Prob}[t_k = 2] \\ &\quad + \text{Prob}[t_{k+1} = 2 | t_k = 1]\text{Prob}[t_k = 1] \\ &= 2\alpha(1-\alpha),\end{aligned}$$

the former reduces to $1/2$:

$$\begin{aligned}\text{Prob}[t_{k+1} \neq t_k | t_k \neq t_{k-1}] &= \frac{\text{Prob}[t_{k+1} \neq t_k,\, t_k \neq t_{k-1}]}{\text{Prob}[t_k \neq t_{k-1}]} \\ &= [\text{Prob}[t_{k+1} = 1 | t_k = 2,\, t_{k-1} = 1]\text{Prob}[t_k = 2,\, t_{k-1} = 1] \\ &\quad + \text{Prob}[t_{k+1} = 2 | t_k = 1,\, t_{k-1} = 2]\text{Prob}[t_k = 1,\, t_{k-1} = 2]] \\ &\quad / \text{Prob}[t_k \neq t_{k-1}] \\ &= \frac{\alpha^2(1-\alpha) + \alpha(1-\alpha)^2}{2\alpha(1-\alpha)} = \frac{1}{2}.\end{aligned}$$

The above probabilities are only equal in case $\alpha = 1/2$, and consequently $\beta = 1/2$, or $\alpha = \beta$, which is exactly the special case that we have dealt with in Section 4.3.

5 Discussion of results and numerical examples

In this section, we revisit the results that were obtained for the general case, both from a qualitative perspective and by means of some numerical examples. In particular, we focus on what we believe is the most commonly occurring case, namely that where the average service time of a customer is longer when the previous customer is of the opposite class (i.e., $\mu_A < \mu_B$). As an example, we recall the two software programs that run on the same server and that both require lots of program-specific data. If program P was executed during the previous run, the necessary data was loaded into the cache memory, and hence execution of program P in the current run will be very fast.

The first interesting result was already given by Equation (5). The equation expresses the direct dependency of the work load ρ ($\triangleq \lambda \mathrm{E}[c]$) on the interclass correlation factor γ. Consequently, the stability condition,

$$\lambda < \frac{1}{\mathrm{E}[c]} = \frac{1}{\mu_A + 2(1-\gamma)(\mu_B - \mu_A)\pi_1\pi_2} \text{ ,} \tag{26}$$

reveals that the supremum of the achievable throughput of the presented system, denoted as λ_{sup} from here on, and expressed in customers per slot, depends on γ. It also depends on the average service time μ_A, the average service-time difference $\mu_B - \mu_A$ and the fractions of class 1 and 2 customers.

Equation (26) reveals that for a fixed average service-time difference $\mu_B - \mu_A$ the achievable throughput of the system gets lower as μ_A increases: longer average service times μ_A and μ_B lead to a lower average number of customers that can be served per time slot. If the difference $\mu_B - \mu_A$ increases (by increasing μ_B), λ_{sup} decreases as well, because the mean service time for customers following customers of the opposite class is increased.

For fixed average service times μ_A and μ_B, we find that λ_{sup} is lowest when $\pi_1\pi_2$ reaches it maximal value, i.e., for $\pi_1 = \pi_2 = \frac{1}{2}$. If one class of customers enters the system more often than the other ($\pi_1 > 0.5 > \pi_2$ or $\pi_2 > 0.5 > \pi_1$), consecutive customers will be of the same class more often, implying that the average service time of an arbitrary customer decreases, or that the throughput of the system increases.

When π_1, π_2, μ_A and μ_B are fixed, the throughput decreases when the classes of consecutive customers alternate more frequently, i.e., when γ becomes smaller. The worst-case scenario occurs when $\gamma = -1$ meaning that the classes of subsequent customers differ all the time. The best case scenario occurs for $\gamma = 1$, when only one class of customers enters the system.

A second interesting result was given by Equation (21). It provides an expression for the average system content at random slot boundaries:

$$\begin{aligned} \mathrm{E}[u] = \rho + \frac{\lambda^2 C''(1) + E''(1)C'(1)}{2(1-\rho)} + 2\lambda(\mu_A - \mu_B)\pi_1\pi_2 + \frac{p_B\lambda(\mu_B - \mu_A)}{1-\gamma} \\ + \frac{(1-\gamma)\lambda^2(\mu_B - \mu_A)^2\pi_1\pi_2(1 - 4\pi_1\pi_2)}{1-\rho} \ . \end{aligned} \tag{27}$$

The expression clearly indicates the influence of the different system parameters on the mean system content at random slot boundaries. The first two terms of Equation (27) correspond to the classical terms that constitute the expression for the average system content at random slot boundaries of a system with no interclass correlation and a service-time pgf $C(z)$. The other three terms in the expression can be fully attributed to the presence of class clustering in the arrival process. Note that this last part depends on the service-time distributions $A(z)$ and $B(z)$ through the difference of their mean values only.

It is not surprising to see that the mean system content depends on the first two moments of the aggregated arrival process (represented by the quantities λ, $E''(1)$ and $\rho = \lambda E[c]$) and on the first two moments of the service times (represent by the quantities $C'(1)$, $C''(1)$, μ_A, μ_B and $\rho = \lambda C'(1)$). Furthermore, as we anticipated, the mean system content goes to infinity as soon as the work load ρ approaches its limiting value 1.

It is also worth noting that it is premature to conclude from Equation (27) that the average system content raises without bound when the interclass correlation factor γ is 1. When $\gamma = 1$, and consequently $\alpha = \beta = 1$, Equation (16) indicates that the numerator of the fourth term of Equation (27) vanishes too, as $p_B = 0$. In fact, when $\gamma = 1$, one resides in a single-class system (see also Section 4.2), implying that the average system content will not rise endlessly as long as the stability condition is obeyed.

In Figures 4-6, we present numerical results for two-class queueing systems dealing with an aggregated Poisson arrival process (i.e., $E(z) = e^{\lambda(z-1)}$) and the following pgfs of the service times for both customers following a customer of the same class and customers following a customer of the opposite class respectively:

$$A(z) = z;\ B(z) = \frac{z}{\mu_B + (1-\mu_B)z} \ . \tag{28}$$

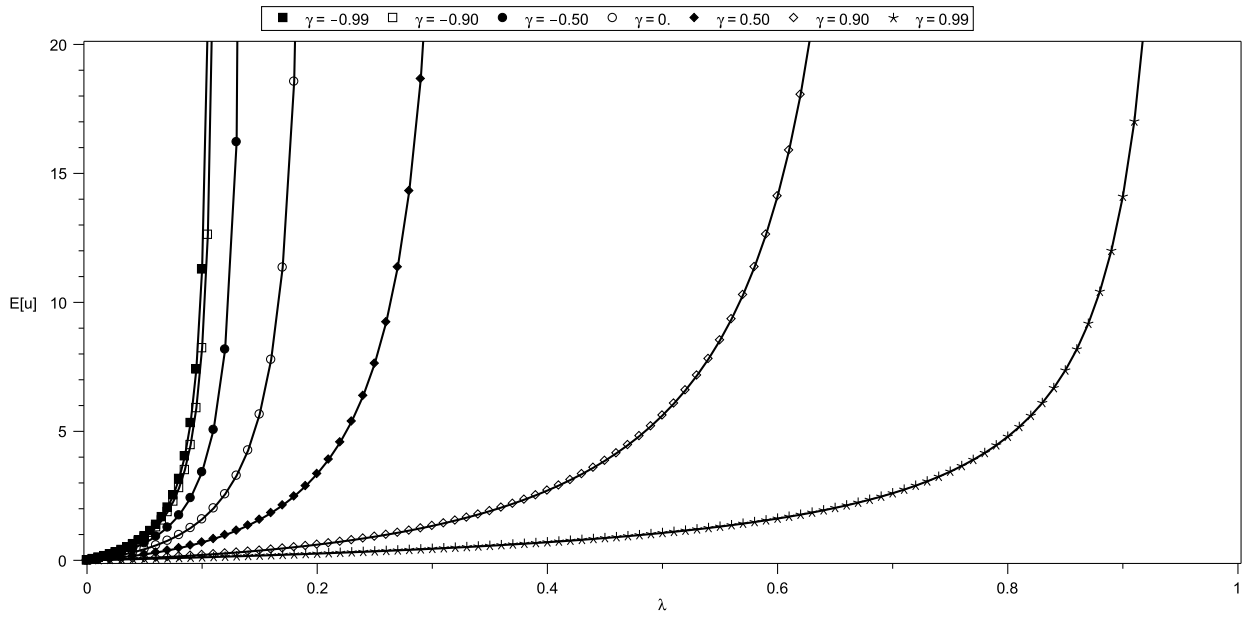

Fig. 4. Mean system content versus the mean arrival rate λ, for Poisson arrivals, $A(z)$ and $B(z)$ given by (28), $\mu_B = 9$, $\pi_1 = \pi_2 = 0.5$ and several interclass correlation factors.

Hence, the service of customers following a customer of the same class only requires one time slot and in the opposite case the service time is geometrically distributed with mean value μ_B. Figure 4 shows the mean system content versus λ for different values of γ, in a system where $\mu_B = 9$ and both classes of customers occur with the same a priori frequency (i.e., $\pi_1 = \pi_2 = 0.5$).

One can observe that the average number of customers each system can deal with depends heavily on the amount of interclass correlation: the more positive that correlation, the more customers can be served per time slot. In terms of average system content this implies that the system occupancy raises rapidly for systems with a negative interclass correlation.

In Figure 5, we examine the impact of the fractions of class 1 and class 2 customers in the arrival stream on the average system content. The figure depicts the mean system content versus λ, in a system where $\mu_B = 9$, and with a fixed interclass correlation of 0.

The figure mainly shows that having two classes of customers instead of one strongly affects the mean system content. If only one class of customer occurs, the average system content is much lower, because every arriving customer only requires one time slot to be served. As soon as two different classes of customers enter the system, the average system content increases considerably. As we reasoned before, based on Equation (26), the exact fraction of class 1 and class 2 customers influences the achievable throughput of the system.

In a third plot (Figure 6), we present the mean system content of a system that is facing a positive interclass correlation of 0.5 and an equal amount of class 1 and class 2 customers. The mean service time μ_B is varied.

As could have been anticipated, we see that the average system content increases when μ_B increases. If the interclass correlation factor is fixed, a longer service time for customers that are not of the same class

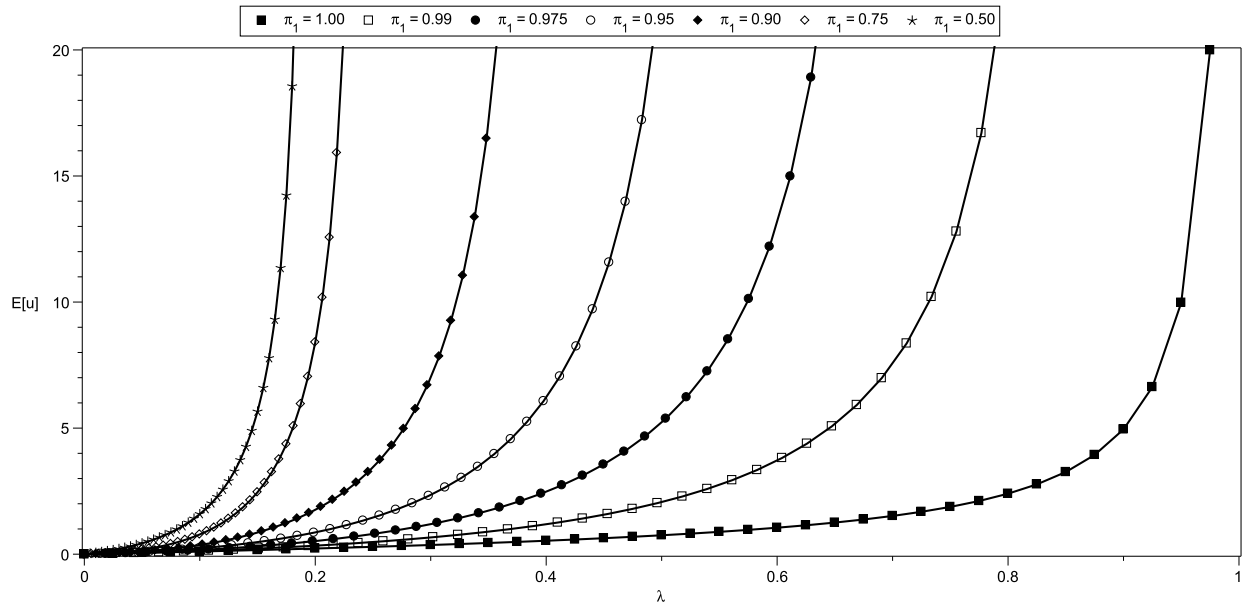

Fig. 5. Mean system content versus the mean arrival rate λ, for Poisson arrivals, $A(z)$ and $B(z)$ given by (28), $\mu_B = 9$, $\gamma = 0$ and various fractions of customer classes in the arrival stream.

as the previous customer implies more arriving customers during that service time, and consequently, more customers waiting in the system to be served.

In order to verify how the modeled systems behave when μ_A is higher than μ_B, we have also examined some examples for that case. This case appears, for instance, when execution of a certain task requires postprocessing (such as cooling down or reinitialisation of a machine). Intuitively, we expected to see the opposite system behaviour as the one that was visualized in Figures 4-6: we anticipated higher average system contents and a smaller stability region for (1) systems with more positive interclass correlation, (2) systems where the fractions of the two customer classes are more balanced and (3) systems where the value of μ_A is increasing. All our expectations were met. To give the reader an example, we present one more plot (Figure 7) in which the mean system content versus λ for different values of γ is depicted, for a system where $\mu_A = 3$ (geometrically distributed service times), $\mu_B = 1$ and $\pi_1 = \pi_2 = 0.5$.

6 Conclusions and future work

In this paper, we have analysed the performance of a system where the service-time distribution of a customer depends on the equality or non-equality of its class with the class of the preceding customer. The major contribution is that we have incorporated class clustering in our model. Although it was intuitively clear that class clustering can have a huge impact, this feature was traditionally overlooked in literature. We have included it in the model, hence providing a more accurate and realistic system analysis. In addition, it enables us to quantify the influence of class clustering. We believe that our model can be more suitable for particular applications, for instance in manufacturing, as compared to traditional multi-class models.

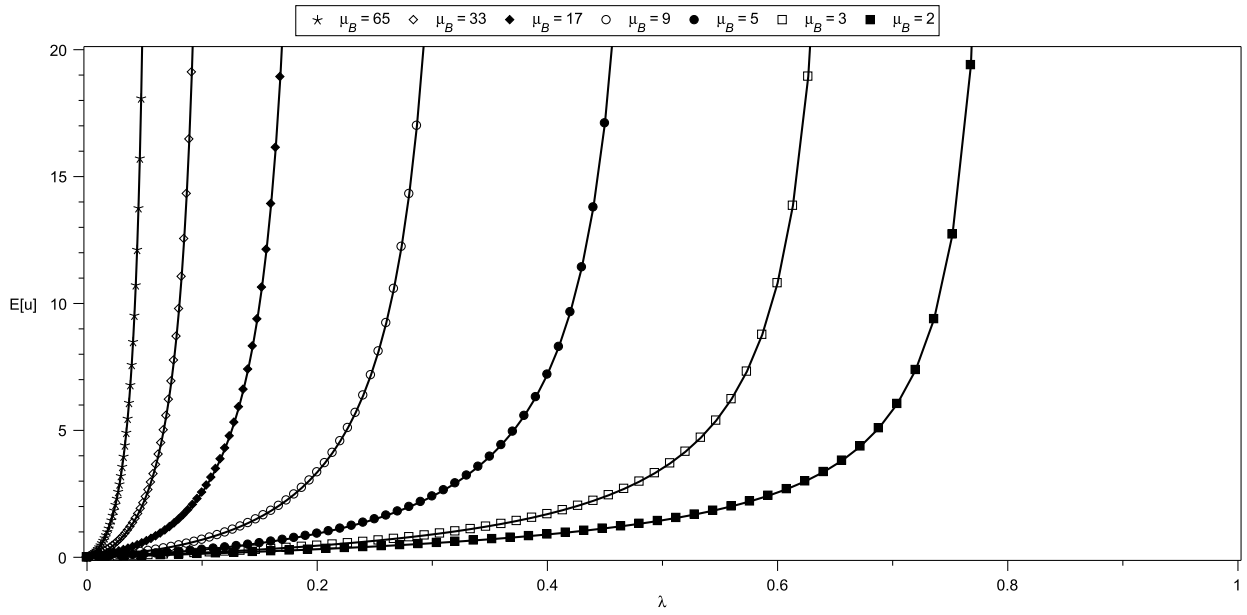

Fig. 6. Mean system content versus the mean arrival rate λ, for Poisson arrivals, $\gamma = 0.5$ and $\pi_1 = \pi_2 = 0.5$; $A(z)$ and $B(z)$ are given by (28).

There are a number of possible extensions to this work. First, the pgf of the customer delay can be of interest. A possible approach could be linking the delay of consecutive customers, as this might make it possible to deal with the class clustering feature. Also, considering more customer classes is an interesting topic for future research. This could require a matrix-analytic solution method, but the amount of numerical work might become considerable. Another possible extension is considering correlation between the total numbers of arrivals during consecutive slots. A possibility might be to model this is a batch-Markovian arrival process. Finally, it would be interesting to compare the results of this paper with results for an analogous model with another scheduling policy than FCFS. The main challenge in that regard is that the class of the last customer that left the system might not provide enough information about the class of the next customer in service, as this customer might not necessarily be the one that arrived after the customer that has just departed.

Acknowledgement

The authors would like to thank the anonymous referees and the editor for their constructive suggestions, which led to a considerable improvement of this paper. This research has been funded by the Interuniversity Attraction Poles Programme initiated by the Belgian Science Policy Office.

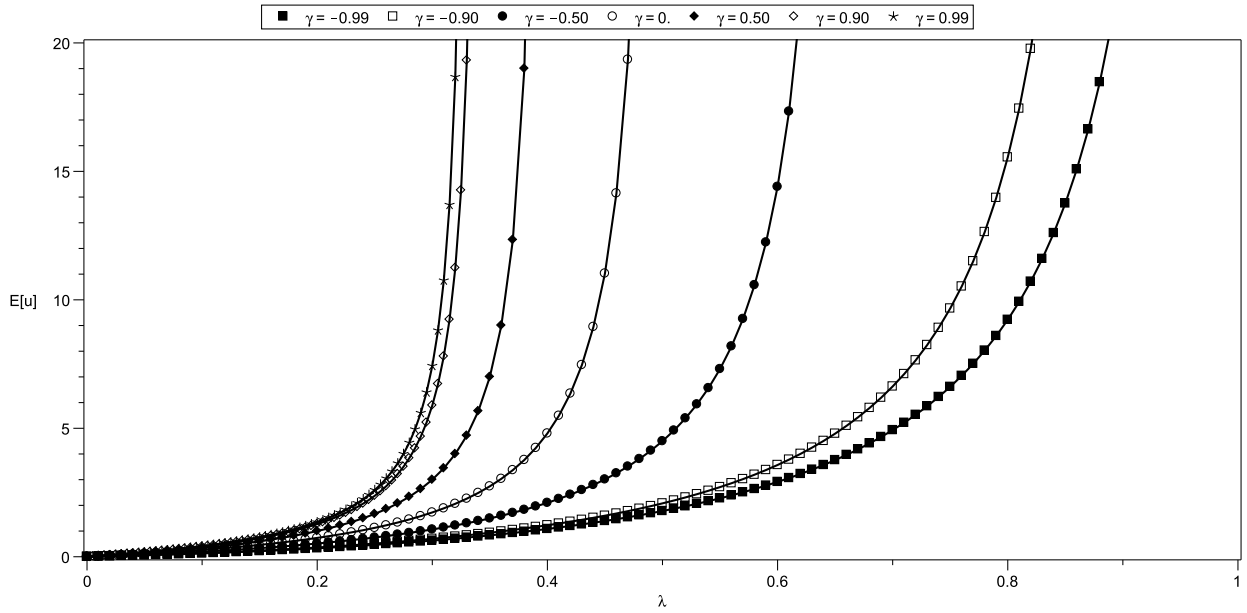

Fig. 7. Mean system content versus the mean arrival rate λ, for Poisson arrivals, $A(z) = \frac{z}{\mu_A+(1-\mu_A)z}$ with $\mu_A = 3$, $B(z) = z$, $\pi_1 = \pi_2 = 0.5$ and several interclass correlation factors.

References

[1] I.J.B.F. Adan, A. Sleptchenko, and G.J. Van Houtum. Reducing costs of spare parts supply systems via static priorities. *Asia-Pacific Journal of Operational Research*, 26(4):559–585, 2009.

[2] I.J.B.F. Adan, J.S.H. van Leeuwaarden, and E.M.M. Winands. On the application of Rouché's theorem in queueing theory. *Operations Research Letters*, 34(3):355–360, 2006.

[3] M.A.A. Boon, I.J.B.F. Adan, and O.J. Boxma. A polling model with multiple priority levels. *Performance Evaluation*, 67:468–484, 2010.

[4] H. Bruneel and B.G. Kim. *Discrete-time models for communication systems including ATM.* Kluwer Academic, Boston, USA, 1993.

[5] H. Bruneel, T. Maertens, B. Steyaert, D. Claeys, D. Fiems, and J. Walraevens. Analysis of a two-class FCFS queueing system with interclass correlation. In *Proceedings of the 19th International Conference on Analytical and Stochastic Modelling Techniques and Applications (ASMTA'12)*, Grenoble, June 4-6 2012.

[6] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys, and J. Walraevens. Impact of blocking when customers of different classes are accommodated in one common queue. In *Proceedings of the 1st International Conference on Operations Research and Enterprise Systems (ICORES)*, Villamoura, Portugal, February 2012.

[7] H. Bruneel, W. Mélange, B. Steyaert, D. Claeys, and J. Walraevens. A two-class discrete-time queueing model with two dedicated servers and global FCFS service discipline. *European Journal of Operational Research*, 223:123–132, 2012.

[8] H. Chen and H. Zhang. Stability of multiclass queueing networks under priority service disciplines. *Operations Research*, 48(1):26–37, 2000.

[9] E.B. Cil, F. Karaesmen, and E.L. Ormeci. Dynamic pricing and scheduling in a multi-class single-server queueing system. *Queueing Systems*, 67(4):305–331, 2011.

[10] D. Claeys, H. Bruneel, B. Steyaert, W. Mélange, and J. Walraevens. Influence of data clustering on in-order multi-core processing systems. *Electronics Letters*, 49(1):28–29, 2013.

[11] W. Feller. *An Introduction to Probability Theory and Its Applications, Vol. 1, Third Edition.* Wiley, New York, 1968.

[12] D. Fiems and H. Bruneel. A note on the discretization of Little's result. *Operations Research Letters*, 30:17–18, 2002.

[13] R.G. Gallager. *Discrete stochastic processes.* Kluwer Academic, Boston/Dordrecht/London, 1996.

[14] D. Gamarnik and D. Katz. On deciding stability of multiclass queueing networks under buffer priority scheduling policies. *Annals of Applied Probability*, 19(5):2008–2037, 2009.

[15] K. Kim and K. Chae. Discrete-time queues with discretionary priorities. *European Journal of Operational Research*, 200(2):473–485, 2010.

[16] M. Larrañaga, U. Ayesta, and I.M. Verloop. Dynamic fluid-based scheduling in a multi-class abandonment queue. *Performance Evaluation*, 70(10):841–858, 2013.

[17] T. Maertens, H. Bruneel, and J. Walraevens. Effect of class clustering on delay differentiation in priority scheduling. *Electronic Letters*, 48(10):568–569, 2012.

[18] W. Mélange, H. Bruneel, B. Steyaert, D. Claeys, and J. Walraevens. Impact of class clustering and global FCFS service discipline on the system occupancy of a two-class queueing model with two dedicated servers. In *Proceedings of the 7th International Conference on Queueing Theory and Network Applications (QTNA 7)*, Kyoto, Japan, 2012.

[19] M. Sidi. Two competing discrete-time queues with priority. *Queueing Systems*, 3:347–362, 1988.
[20] M. Sidi and A. Segall. Structured priority queueing systems with applications to packet-radio networks. *Performance Evaluation*, 3(4):265–275, 1983.
[21] H. Takagi. *Queueing analysis - vol. 3: discrete-time systems.* North Holland, 1993.
[22] O.S. Uluscu and T. Altiok. Waiting time approximation in multi-class queueing systems with multiple types of class-dependent interruptions. *Annals of Operations Research*, 202(1):185–195, 2013.
[23] I.M. Verloop, U. Ayesta, and S. Borst. Monotonicity properties for multi-class queueing systems. *Discrete Event Dynamic Systems - Theory and Applications*, 20(4):473–509, 2010.
[24] J. Walraevens, D. Fiems, and H. Bruneel. Time-dependent performance analysis of a discrete-time priority queue. *Performance Evaluation*, 65:641–652, 2008.